AFRL-OSR-VA-TR-2014-0234

# STATISTICAL MACHINE LEARNING FOR STRUCTURED AND HIGH DIMENSIONAL DATA

Larry Wasserman
CARNEGIE MELLON UNIVERSITY

09/17/2014
Final Report

DISTRIBUTION A: Distribution approved for public release.

Air Force Research Laboratory
AF Office Of Scientific Research (AFOSR)/ RTC
Arlington, Virginia 22203
Air Force Materiel Command

# REPORT DOCUMENTATION PAGE

*Form Approved*
*OMB No. 0704-0188*

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**

**1. REPORT DATE** *(DD-MM-YYYY)*
14-06-2014

**2. REPORT TYPE**
Final

**3. DATES COVERED** *(From - To)*
Dec 2009 - Aug 2014

**4. TITLE AND SUBTITLE**
Statistical Machine Learning for Structured and High Dimensional Data

**5a. CONTRACT NUMBER**
FA9550-09-1-0373

**5b. GRANT NUMBER**

**5c. PROGRAM ELEMENT NUMBER**

**6. AUTHOR(S)**
Wasserman, Larry
Lafferty, John

**5d. PROJECT NUMBER**

**5e. TASK NUMBER**

**5f. WORK UNIT NUMBER**

**7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)**
Carnegie Mellon University
5000 Forbes Avenue
Pittsburgh, PA 15213

University of Chicago
5801 S. Ellis Avenue
Chicago, IL 60637

**8. PERFORMING ORGANIZATION REPORT NUMBER**

**9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)**
AFOSR: Information, Decision, and Complex Networks
Wright-Patterson Airforce Base

**10. SPONSOR/MONITOR'S ACRONYM(S)**
AFOSR/RTC

**11. SPONSOR/MONITOR'S REPORT NUMBER(S)**

**12. DISTRIBUTION / AVAILABILITY STATEMENT**

**13. SUPPLEMENTARY NOTES**

**14. ABSTRACT**
Research under this grant was carried out in the general area of nonparametric statistical modeling of high dimensional and structured data. The project made a number of advances in methodology and supporting theory for estimating high dimensional regression functions, classification functions, graphical models, and probability densities. Advances were also made on the new research area of resource-constrained statistical estimation.

**15. SUBJECT TERMS**
machine learning, high-dimensional statistics

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| **a. REPORT** | **b. ABSTRACT** | **c. THIS PAGE** | | | John Lafferty |
| U | U | U | UU | | **19b. TELEPHONE NUMBER** *(include area code)* 773-702-3813 |

**Standard Form 298 (Rev. 8-98)**
**Prescribed by ANSI Std. Z39.18**

>

Research under this grant was carried out in the general area of nonparametric modeling of high dimensional and structured data. We made a number of advances, briefly summarized below.

The Nonparanormal SKEPTIC

In this work we develop a semiparametric approach to efficiently and robustly estimating high dimensional undirected graphical models. To achieve modeling flexibility, we consider Gaussian Copula graphical models (or the nonparanormal). To achieve estimation robustness, we exploit nonparametric rank-based correlation coefficient estimators, including Spearman's rho and Kendall's tau. In high dimensional settings, we prove that the nonparanormal SKEPTIC achieves the optimal parametric rate of convergence in both graph and parameter estimation.

Sparse Nonparametric Graphical Models

The Gaussian graphical model is the standard parametric model for continuous data, but it makes distributional assumptions that are often unrealistic. We developed two approaches to building more flexible graphical models. One allows arbitrary graphs and a nonparametric extension of the Gaussian; the other uses kernel density estimation and restricts the graphs to trees and forests.

Sparse Additive Functional and Kernel CCA

Canonical Correlation Analysis (CCA) is a classical tool for finding correlations among the components of two random vectors. In recent years, CCA has been widely applied to the analysis of genomic data, where it is common for researchers to perform multiple assays on a single set of patient samples. Recent work has proposed sparse variants of CCA to address the high dimensionality of such data. However, classical and sparse CCA are based on linear models, and are thus limited in their ability to find general correlations. We developed two approaches to high-dimensional nonparametric CCA, building on recent developments in high-dimensional nonparametric regression.

Sequential Nonparametric Regression

We developed algorithms for nonparametric regression in settings where the data are obtained sequentially. While traditional estimators select bandwidths that depend upon the sample size, for sequential data the effective sample size is dynamically changing. We developed a linear time algorithm that adjusts the bandwidth for each new data point, and showed that the estimator achieves the optimal minimax rate of convergence.

Computation-risk tradeoffs for covariance-thresholded linear regression.

Modern data sets used for statistical analysis are often large and high dimensional. The computation required to construct standard estimators for such data may be prohibitive. In this setting it is attractive to tradeoff statistical accuracy for computational scalability--tolerating increased predictive error, or risk, in exchange for more favorable computational requirements. But little is known about precise tradeoffs between risk and computation.

We have studied such tradeoffs in the setting of large scale linear regression, developing a concrete, practical way to smoothly tradeoff risk for computation, by sparsifying the sample covariance with hard thresholding. The main of this work is to combine recent computational developments for sparse symmetric diagonally dominant linear systems with a statistical analysis of the predictive risk for this family of linear models, making precise the tradeoff between computation and error.

Computation-risk tradeoffs in nonparametric regression using locality-sensitive hashing

Related to our work on linear regression, we have developed a new approach to trading off statistical risk for speed of computation in nonparametric regression. We adopt the classical kernel smoothing estimator to incorporate variants of locality-sensitive hashing to quickly find points in a neighborhood of the query point. The maximum number of points that is returned by the hashing scheme acts as a computational tuning parameter. The procedure adapts to the case where the data lie on a low-dimensional manifold. We analyzed the predictive risk of the procedure as a function of the bandwidth and tuning parameters of the hashing algorithm. This demonstrates a setting where the speed of computing the function estimate can be traded off against its accuracy in a fine-grained manner.

Graphical Exponential Screening

We have developed a new mixture estimator of the inverse covariance matrix of a Gaussian graphical model. The estimator, called \graphical Exponential Screening (gES), linearly combines estimators from various models with different underlying graphs and can adaptively balance the mean squared error and sparsity. We prove an oracle inequality for this mixture estimator, showing that it is comparable or even superior to the risk of the best estimator based on a single graph. A key tool in our analysis is an unbiased estimate of the risk of the gES estimator, which generalizes Stein's unbiased risk estimate (SURE) to Wishart distributions. The resulting estimator is free of any tuning parameters, and enjoys strong theoretical properties.

Localized minimax risk for stochastic convex optimization

Traditional minimax theory gives a worst cast analysis of statistical estimation procedures. This worst-case approach is also the standard for the analysis of algorithms. We have been studying a ``softening'' of this traditional worst-case analysis called local minimax complexity. In local minimax complexity, we consider the ``hardest local alternative'' to minimizing a specific function. We have shown that local minimax complexity for stochastic optimization can be bounded in terms of the modulus of continuity, and gives tight bounds for optimizing convex functions in many cases. This ties together the statistical and computational perspectives, and leads to a more relevant theoretical analysis for computational problems.

Quantization-risk tradeoffs in statistical estimation

We have studied quantization-risk tradeoffs for nonparametric estimation. Here we consider bounds on the number of bits used to express an estimator. Our approach combines elements of rate-distortion theory and statistical minimax theory. We have analyzed the case of the normal means within a Sobolev ellipsoid, which is a standard setup in nonparametric regression. Our results show how the excess risk varies with the number of bits used to represent the estimate. Our thinking on this problem was motivated by large scale data analysis problems. In particular, we have been working with data from the Kepler telescope for finding exoplanets orbiting distant stars. The telescope cannot send the raw data back to earth directly because of communication constraints. Instead, it averages and subsamples. Our theory determines the least possible increase in risk that results from the best B-bit representation of the data.

Convex regression

Building on our earlier work on sparse additive models, we have studied the problem of estimating a sparse convex function of many variables. In contrast to classical nonparametric regression with smoothness constraints, we show that convexity is additively faithful--it suffices to estimate a convex additive model for variable selection. We develop algorithms for estimating sparse convex additive models, including an approach using iterative quadratic programming. Supporting experiments and statistical theory together show that this approach achieves variable selection consistency in dimensions that can scale exponentially in the sample size. An attractive feature of this framework is the lack of tuning parameters for smoothness.

Publications

The Nonparanormal SKEPTIC
Han Liu, Fang Han, Ming Yuan, John Lafferty, Larry Wasserman
ICML 2012.

Huge: High Dimensional Undirected Graph Estimation Tuo Zhao, Han Liu, Kathryn Roeder, John Lafferty, Larry Wasserman Journal Of Machine Learning Research (JMLR) Vol 13, 1059-1062, 2012

Nonparametric reduced rank regression
Rina Foygel, Michael Horrell, Mathias Drton, and John Lafferty
Advances in Neural Information Processing Systems 25, pp 1637-1645, 2012.

Exponential concentration for mutual information estimation with application to forests
Han Liu, John Lafferty, Larry Wasserman
Advances in Neural Information Processing Systems 25, pp 2546-2554, 2012.

High-dimensional semiparametric Gaussian copula graphical models
Han Liu, Fang Han, Ming Yuan, John D. Lafferty and Larry A. Wasserman
Ann. Statist. Volume 40, Number 4, 2012, 2293-2326.

Sparse nonparametric graphical models
John Lafferty, Han Liu, and Larry Wasserman
Statist. Sci. Volume 27, Number 4, 2012, 519-537.

Sequential nonparametric regression Haijie Gu and John D. Lafferty, Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland.

Computation-risk tradeoffs for covariance-thresholded regression, Dinah Shender and John Lafferty, Proceedings of the International Conference on Machine Learning (ICML-13), Volume 28-3, pp. 756-764.